

Differential Item Functioning: Current Issues

Gi-Pyo Park
(Soonchunhyang University)

Park, Gi-Pyo. (2005). Differential item functioning: Current issues. *Language Research* 41(4), 949-962.

This study describes current issues of differential item functioning (DIF) used by psychometricians and test specialists to identify item fairness across members of different subgroups such as males and females. First, the definition of DIF was discussed, centering on Simpson's paradox and two different types of DIF: uniform DIF and non-uniform DIF. Then, three commonly used DIF detection techniques – the Mantel-Haenszel method, confirmatory factor analysis (CFA), and the likelihood ratio (LR) test – were introduced. Third, the relation of DIF to differential test functioning (DTF) was explored with the generation of three research hypotheses based on current empirical studies. Finally, this study was concluded with future research inquiries to be answered.

Key words: test bias, differential item functioning, differential test functioning

1. Introduction

Since a test result has personal, social, and political ramifications, it should be reliable, valid, and fair.¹⁾ In order to investigate whether a test item is fair among members of different subgroups such as males and females and majority groups and minority groups, a plethora of research on differential item functioning (DIF) has been conducted (Elder, 1997; Holland & Wainer, 1993; Maller, 2003; Ryan & Bachman, 1992).

Previous research on DIF has produced important findings in terms of definition and DIF detection techniques. However, whether a test with DIF items manifests different test functioning (DTF) is not fully explored yet. Research on the relation of DIF to DTF is critical because it provides relevance of DIF studies and because decisions with a test are made by

1) This work was supported by Soonchunhyang University Research Grant 20050055.

the results of a whole test score (Roznowski & Reith, 1999).

To date, a total of six empirical studies have been undertaken to examine whether items showing DIF manifest DTF on a test level analysis (Drasgow, 1987; Pae & Park, in press; Roznowski, 1987; Roznowski & Reith, 1999; Takala & Kaftandjieva, 2000; Zumbo, 2003). For this, these studies have used such DIF detection techniques as the Mantel-Hanszel procedure, confirmatory factor analysis (CFA), and the likelihood ratio (LR) test among the many techniques developed so far (Maller, 2003; Millsap & Everson, 1993).

The major purpose of this study was to generate research hypotheses regarding the relation of DIF to DTF by reviewing previous empirical studies. Other supplementary purposes of this study were to clarify the definition of DIF centering on group comparability and two different types of DIF and to delineate three aforementioned DIF detection techniques. Finally, this study was intended to sensitize DIF and DTF developed by psychometricians, but published specifically in *Language Testing*, to L2 acquisition researchers.

2. Definition of DIF

Differential item functioning (DIF) is present when two groups of equal ability show a differential probability of a correct response to an item (Ellis & Raju, 2003). It should be noted that DIF is different from item bias and item impact (Angoff, 1993). DIF is determined by the “simple observation” of different statistical properties across groups, whereas bias is determined by the “thoughtful judgment” of different statistical properties for members of different subgroups. Item impact refers to group differences due to the differences of true group ability rather than group favoritism.

In the definition of DIF “two groups of equal ability” is essential because DIF may exist beyond mean differences between two groups. More specifically, Simpson’s paradox illustrates why groups of equal ability should be compared (Dorans & Holland, 1993; see also Thissen et al., 1986). Table 1 summarizes the performance of two groups – the focal group (a group of primary interest) and the reference group (a standard group against which the focal group is compared) – for an imaginary item. N_m , N_{cm} , and N_{cm}/N_m in each group refer to people at the ability level m ,

people who answered the item correctly at the ability level m , and the proportion of people who answered the item correctly at the ability level m , respectively. The rows in each group refer to the levels from lower to higher, with the fourth row indicating the sum of each ability level.

Table 1. Performance of Two Groups on an Imaginary Item

Focal Group			Reference Group		
N_m	N_{cm}	N_{cm}/N_m	N_m	N_{cm}	N_{cm}/N_m
400	40	.10	1000	200	.20
1000	500	.50	1000	600	.60
1000	900	.90	400	400	1.00
2400	1440	.60	2400	1200	.50

When the examinees at each ability level are compared together, the item favors the focal group by .10 ($0.60 \sim 0.50$). However, when the examinees at each ability level are compared separately, the item is in favor of the reference group at all the levels by .10 ($.20 \sim .10$, $.60 \sim .50$, $1.00 \sim .90$). This contradiction of group favoritism is due to unequal distributions of ability between the two groups in the N_m . Thus, Table 1 evidences the essence of two groups of equal ability or after matching the ability between two groups in the definition of DIF.

Two types of DIF have been discussed in the literature: uniform DIF and non-uniform DIF. Uniform DIF is present when an item differs across groups in item difficulty parameters, while non-uniform DIF is present when an item differs across members of different subgroups in item discrimination parameters. The difference between uniform DIF and non-uniform DIF is shown in Figure 1 and in Figure 2 (Clauser & Mazor, 1998).

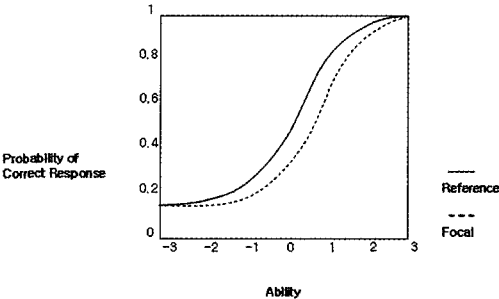


Figure 1. Uniform DIF

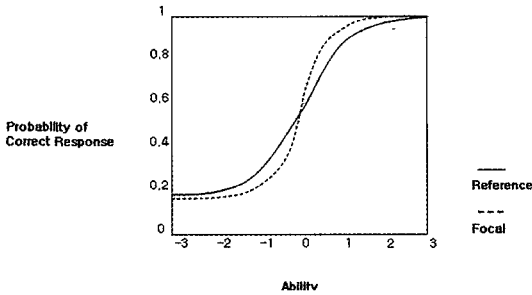


Figure 2. Non-uniform DIF

Uniform DIF and non-uniform DIF in Figures 1 and 2 are illustrated by item response theory (IRT). In Figure 1, the item is in favor of the reference group across ability levels. In Figure 2, however, the item favors either the reference group ($\theta < 0$) or the focal group ($\theta > 0$), respectively, depending on the ability levels (θ) (Hambleton et al., 1991).

3. DIF Detection Techniques

Many techniques have been developed to examine measurement invariance such as the Mantel-Haenszel procedure (Holland & Thayer, 1988), standardization approach (Dorans & Holland, 1993), logistic regression method (Swaminathan & Rogers, 1990), the likelihood ratio test (Thissen et al., 1993), Lord's chi-square test (Lord, 1980), and confirmatory factor analysis (Jöreskog, 1971). Among these techniques, the Mantel-Haenszel procedure and confirmatory factor analysis (CFA) based on classical test theory and the likelihood ratio (LR) test based on item response theory have been popularly used, specifically in the studies examining the relation of DIF to DTF.

3.1. The Mantel-Haenszel Procedure

Holland and Thayer (1988) adapted the Mantel-Haenszel procedure (Mantel & Haenszel, 1959) to identify items displaying DIF. The Mantel-Haenszel procedure uses the $2 \times 2 \times M$ contingency table in which an item is arranged in two groups (the focal group and the reference group), two levels of response (right or wrong), and M score levels, as seen in Table 2.

Table 2. The $2 \times 2 \times M$ Contingency Table

		Item Score		
		Right	Wrong	Total
Group	Focal Group	R_{fm}	W_{fm}	N_{fm}
	Reference Group	R_{rm}	W_{rm}	N_{rm}
	Total Group	R_{tm}	W_{tm}	N_{tm}

If there is no DIF, the odds of getting an item correct at a given ability level is the same both in the focal group and in the reference group, and can be expressed as the equation (1):

$$(1) H_0: [R_{rm}/W_{rm}]/[R_{fm}/W_{fm}] = 1 \quad m = 1, 2, \dots, M$$

Using the common odds ratio α , the equation (1) can be rewritten as the equation (2). The parameter α is called the common odds ratio because under H_a , the value of α is the odds ratio that is the same for all m in the $2 \times 2 \times M$ contingency table. There is no DIF when $\alpha = 1$, whereas there is DIF when $\alpha \neq 1$.

$$(2) H_a: [R_{rm}/W_{rm}] = \alpha [R_{fm}/W_{fm}] \quad m = 1, 2, \dots, M \text{ and } \alpha = 1 \text{ or } \alpha \neq 1$$

The null hypothesis in the equation (2), $H_0: \alpha = 1$, can be tested using the Mantel-Haenszel chi-square statistic:

$$(3) MH-\chi^2 = \frac{[\sum R_{rm} - \sum E(R_{rm}) - 0.5]^2}{\sum \text{var}(R_{rm})}$$

where $E(R_{rm}) = N_{rm}R_{tm}/N_{tm}$ and

$$\sum \text{var}(R_{rm}) = [N_{rm}R_{tm}N_{fm}W_{tm}]/[N_{tm}^2(N_{tm} - 1)].$$

Mantel and Haenszel (1959) provided an estimate of the constant odds ratio (α_{MH}) which can be converted into a difference in deltas (Δ) as in the equations (4) and (5), respectively.

$$(4) \alpha_{MH} = [\sum_m R_{rm} W_{fm} / N_{tm}] / [\sum_m R_{fm} W_{rm} / N_{tm}]$$

$$(5) \Delta_{MH} = -2.35 \ln[\alpha_{MH}]$$

The value of delta is ranged from $-\infty$ to $+\infty$. ETS suggested negligible DIF if the absolute value of $\Delta_{MH} < 1.0$, intermediate DIF if the absolute

value of $1.0 \leq \Delta_{MH} < 1.5$, and large DIF if the absolute value of $\Delta_{MH} \geq 1.5$.

In short, the measurement invariance in the Mantel-Haenszel procedure is to test $H_0: \alpha = 1$. Compared with CFA and the LR test which use an underlying latent trait to estimate ability, the Mantel-Haenszel procedure uses observed scores as a matching criterion. It is important to note that the Mantel-Haenszel procedure performs DIF screening to remove DIF items in total scores.

2. Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) has been used to examine measurement invariance across members of different groups by accounting for the covariance between test items (Jöreskog, 1971; Sörbom, 1974; Byrne et al., 1989; Reise et al., 1993; Flowers et al., 2002). Reise et al. (1993) schematized the relationship between observed variables and underlying constructs through CFA as:

$$(6) X_m = \lambda_{mp} \varepsilon_p + \delta_m$$

where X_m is a measured variable, λ_{mp} is a regression coefficient, ε_p is a linear function of a latent variable, δ_m is an error term, $m = 1, \dots, n$, and $p = 1, \dots, r$. Assuming n measured variables and r latent variables, the equation (6) can be rewritten as the equation (7):

$$(7) \Sigma_g = \Lambda_g \Phi_g \Lambda_g' + \Psi_g = S_g$$

where Σ is the $(n \times n)$ population covariance matrix among the measured variables, Λ is a $(n \times r)$ matrix of loadings of the n measured variables on the r latent variables, Φ is a $(r \times r)$ matrix of covariances among the latent variables, Ψ is a $(n \times n)$ matrix of covariances among the residuals, S is a sample covariance matrix, and g is the g th sample. It should be noted that the test of measurement invariance through multi-group CFA is to investigate whether the factor loading matrix Λ_g is invariant across groups. That is, the measurement invariance in CFA is to test $H_0: \Lambda_1 = \Lambda_2$ by examining the chi-square of the general model and the restrictive model across groups. In the general model, the values of Λ_g , Φ_g , and Ψ_g matrices for each S_g in the equation (7) are freely estimated, whereas Λ_g

matrix for each S_g is constrained to match ability between groups in the restrictive model. The fit of the model is assessed using the fit statistics of goodness-of-fit index (GFI), the comparative fit index (CFI), and the root mean square residual (RMSEA). These fit indices are interpreted as the proportion of the observed variances and covariances, the difference in the fit of the null and target models, and the fit of the empirical and modeled variance-covariance matrices, respectively (Maller, 2003).

3. The Likelihood Ratio Test

The likelihood ratio (LR) test identifies DIF by accounting for item responses across the focal and reference groups based on IRT which contains one parameter, two parameter, and three parameter logistic models. For instance, the probability of correct response to an item in the three parameter logistic (PL) model includes the item difficulty, item discrimination, and guessing parameters as seen in the equation (8):

$$(8) \quad P(x = 1 | \theta) = C + \frac{1-c}{1+e^{-Da(\theta-b)}}$$

where x is an item response, θ is the estimated ability, a is the item discrimination parameter, b is the item difficulty parameter, c is the pseudo-guessing parameter, D is a scaling factor to make the logistic function close to the normal ogive function, and e is a transcendental value of 2.718 (Hambleton et al., 1991).

The measurement invariance in the LR test is to test $H_0: a_1 = a_2$ and $H_0: b_1 = b_2$ by examining the likelihood of an augmented model to that of a compact model as seen in the equation (9):

$$(9) \quad G^2(d.f.) = 2 \log \left[\frac{\text{Likelihood}(A)}{\text{Likelihood}(C)} \right]$$

where Likelihood represents likelihood of the data given the maximum likelihood estimates of the parameters of the model, df is the difference between the number of parameters in the augmented model and the number of parameters in the compact model, A is the augmented mod-

el, and C is the compact model (Thissen et al., 1986; Thissen et al., 1993).

In the compact model, all item parameters of the focal and reference groups were constrained equally to match ability between the groups, whereas equality constraints were not imposed in the augmented model. Matching ability across groups can be performed by using purified or anchor items displaying no DIF across groups and can be found by other DIF detection methods such as the Mantel-Haenszel procedure. Then, as described in the Mantel-Haenszel procedure, each studied item is added to anchor items and tested for DIF (Maller, 2001).

4. Relation of DIF to DTF

One of the essential areas in the studies of DIF is to investigate whether DIF leads to differential test functioning (DTF). A test shows DTF if two groups of equal ability show a differential probability of a correct response to a test as a whole (Ellis & Raju, 2003). Based on the empirical studies, the relation of DIF to DTF can be hypothesized as follows:

Hypothesis 1: A test with DIF items shows DTF because the amount of DIF is accumulated in the whole test level analysis.

Hypothesis 2: A test with DIF items shows no DTF because DIF items cancel out each other in the whole test level analysis.

Hypothesis 3: A test with DIF items shows no DTF because DIF items are not detrimental in the whole test level analysis independent of DIF cancellation.

Several empirical studies investigating the effect of remaining DIF items on the whole test by using such methods as test characteristic curves, factor structure invariance, and prediction of criteria have been published specifically in *Language Testing*. These studies have supported either the Hypothesis 1 (Pae & Park, in press), the Hypothesis 2 (Drasgow, 1987; Ryan & Bachman, 1992; Takala & Kaftandjieva, 2000), or the Hypothesis 3 (Roznowski, 1987; Roznowski & Reith, 1999; Zumbo, 2003).

Research supporting the Hypothesis 1 was undertaken by Pae and Park (in press). He identified DIF with the LR test in the English reading test of 2003 college scholastic aptitude test (CSAT), and the relation of

DIF to DTF was investigated with multigroup CFA. A total of 22 DIF items in b (item difficulty) parameters were found in the 33 items of the reading test, with 13 items in favor of males and 9 items in favor of females. Interestingly, the effect of DIF was manifest in the test level analysis without showing DIF cancellation, as measured by factor loadings, factor variances and covariances, and error variances.

Research supporting the Hypothesis 2 includes the studies by Drasgow (1987), Ryan and Bachman (1992), and Takala and Kaftandjieva (2000). Drasgow (1987) identified several DIF items across gender and race in the ACT Mathematics Usage test with IRT. However, test-characteristic curves, which are the sum of the item-characteristic curves and show the net effect of item bias through expected numbers of right scores, identified no group difference in the cumulative effects of DIF items in the test as a whole. Drasgow argued that items with DIF did not show DTF probably because they cancelled out each other in the whole test level.

Ryan and Bachman (1992) detected DIF in the test of English as a foreign language (TOEFL) and in the first certificate of English (FCE) across gender and language background (Indoeuropean/Non-indoeuropean) with the Mantel-Haenszel procedure. In terms of gender, 4 and 2 items out of the 140 TOEFL items were in favor of males and females, respectively, and 1 item out of the 38 FCE items was in favor of both males and females, respectively. For language background, 32 and 33 items in TOEFL were differentially easier for males and females, respectively, and 13 and 12 items in FCE were differentially easier for males and females, respectively. Even though Ryan and Bachman did not mention the relation of DIF to DTF, this study supports Hypothesis 2 because DIF items favoring each group could cancel out each other.

Takala and Kaftandjieva (2000) examined DIF in the vocabulary subtest of the Finnish Foreign Language Certificate Examination with IRT. A total of 11 DIF items were found in the test, with 6 items in favor of males and 5 items favoring females. Despite these DIF items, however, excluding the items with DIF did not make a large difference from the total items in the ability parameter estimations between different subgroups of males and females. These findings implied that the vocabulary test as a whole showed no DTF because the items with DIF balanced out each other in the test level.

The studies by Roznowski (1987), Roznowski and Reith (1999), and

Zumbo (2003) belong to the category of the Hypothesis 3. Roznowski (1987) examined the effects of including items of non-trait variance (possible DIF items) on a test level quality by correlating two subtests taken from the Project TALENT testing battery (each subtest has sex-advantage composite) and the Project TALENT intelligence composite. They found that the correlation between the subtests favoring each gender and the criterion increased when the subtests were combined together. These findings showed that including items of non-trait variance could upgrade measurement characteristics in the whole test level.

Roznowski and Reith (1999) detected DIF items with the Mantel-Haenszel procedure across gender and race and investigated further whether retaining these items was detrimental to the whole test. After creating various item composites such as no bias, both bias, focal bias, and referential bias based on the common odds ratio (α), they investigated the coefficient alphas of the composites followed by correlation and regression analyses. Results showed that the quality of the items with DIF was as good, if not better than, as those with non-DIF.

Zumbo (2003) investigated whether DIF items identified with IRT manifested themselves in a test level analysis conducted with multi-group CFA. Zumbo controlled DIF items in terms of number from 1, 4, 8, to 16 out of 38 DIF items and level from moderate to large through artificial data simulated by the structure and written expression of TOEFL. Results showed that items showing DIF regardless of their number and level did not affect test level invariance.

In sum, in the studies, to date, item level DIF has led more to non-DTF than to DTF. The underlying reasons for this result may be because DIF items cancel out each other in the test level analysis as seen in the Hypothesis 2 or because DIF items are not deteriorating a test as a whole as seen in the Hypothesis 3.

5. Conclusion

This study described up-to-date issues on DIF, centering on the definition of DIF, techniques used to detect DIF, and the relation of DIF to DTF. Indeed, great strides have been made in the studies on DIF for the last 20 years. Nevertheless more studies should be undertaken for a better understanding and application of DIF studies.

The implication of this study to language testing is that the items in an item bank should be pretested for any problems in psychometric properties including DIF and DTF before they are used. If any item shows DIF, the item should be revised or eliminated after thoughtful evaluation by experts. In case any item can not be pretested for security reasons like the items in the Korean College Scholastic Aptitude Test, the selection committee should carefully choose items free from DIF across subgroups such as gender, academic backgrounds, and socio-economic status.

This study leads to the following future inquiries to be answered: First, three research hypotheses raised in this study regarding whether cumulative DIF items lead to DTF should be tested further using a variety of methods such as test characteristic curves, factor structure invariance, and prediction of criterions. Testing these hypotheses is essential because it will provide the relevance of DIF studies (Ryan & Bachman, 1992). Another important reason is that decisions with a test are not made on an individual item level, but on a whole test score level.

Second, the sources of DIF should be identified. Current studies on DIF have successfully identified both uniform and non-uniform DIF. A logical next concern is to identify the sources of DIF after thoughtful judgment of statistical properties by experts. The problem is that identifying the sources of DIF is by no means an easy task even for sensitivity experts (Engelhard, 1990; Elder, 1997). This may be because each item is affected by many variables such as linguistic knowledge, background knowledge, and test-taking strategies (Park, 2004).

In short, testing the research hypotheses generated in this study and identifying the sources of DIF are imminent challenges lying ahead in the future studies of DIF.

References

- Angoff, W. H. (1993) Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NC: Lawrence Erlbaum.
- Byrne, B., Shavelson, R., & Muthn, B. (1989) Testing for the equivalence of factor covariance and mean structures: The issue of partial meas-

- urement invariance. *Psychological Bulletin*, 105, 456-66.
- Clauser, B. E. and Mazor, K. M. (1998) Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17, 31-47.
- Dorans, N. and Holland, P. (1993) DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NC: Lawrence Erlbaum.
- Drasgow, F. (1987) Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72, 19-29.
- Elder, C. (1997) What does test bias have to do with fairness? *Language Testing*, 14, 261-277.
- Ellis, B. and Raju, N. (2003) Test and item bias: What they are, what they aren't, and how to detect them. ERIC ED 480042: Educational Resources Information Center.
- Engelhard, G. (1990) Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education*, 3, 347-360.
- Flowers, C., Raju, M., and Oshima, T. (2002) A comparison of measurement equivalence methods based on confirmatory factor analysis and item response theory. ERIC ED 463302: Educational Resources Information Center.
- Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991) *Fundamentals of item response theory*. Thousand Oaks, CA: Sage.
- Holland, P. and Thayer, D. (1988) Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.
- Holland, P. and Wainer, H. (Eds.) (1993) *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Jreskog, K. G. (1971) Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409-26.
- Lord, F. (1980) Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.
- Maller, S. (2001) Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement*, 61, 793-817.
- Maller, S. (2003) Best practices in detecting bias in nonverbal test. In R. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 23-47).

Kluwer Academic/Plenum Publishers.

- Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from the retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Millsap, R. and Everson, H. (1993) Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17, 297-334.
- Pae, T.-I. and Park, G.-P. (in press). Examining the relationship between differential item and test functioning. *Language Testing*.
- Park, G.-P. (2004) Comparison of L2 listening and reading comprehension by university students learning English in Korea. *Foreign Language Annals*, 37, 448-458.
- Reise, S., Widaman, K., and Pugh, R. (1993) Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114, 552-66.
- Roznowski, M. (1987) Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology*, 72, 480-83.
- Roznowski, M. and Reith, J. (1999) Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement*, 59, 248-69.
- Ryan, K. and Bachman, L. (1992) Differential item functioning on two tests of EFL proficiency. *Language Testing*, 9, 12-29.
- Srbom, D. 1974: A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, 27, 229-39.
- Swaminathan, H. and Rogers, H. (1990) Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- Takala, S. and Kaftandjieva, F. (2000) Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, 17, 323-40.
- Thissen, D., Steinberg, L., and Gerrard, M. (1986) Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, 99, 118-128.
- Thissen, D., Steinberg, L., and Wainer, H. (1993) Detection of differential item functioning using the parameters of item response models. In Holland, P.W. and Wainer, H., editors, *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 35-66.

Zumbo, B. (2003) Does item-level DIF manifest itself in scale-level analysis? Implications for translating language tests. *Language Testing*, 20, 136-47.

Gi-Pyo Park
Department of English
Soonchunhyang University
646 Eupnae-ri, Shinchang-myun, Asan
Choongchungnam-do, Korea 336-745
E-mail: gipyop@sch.ac.kr

Received: Nov. 2, 2005

Revised version received: Dec. 6, 2005

Accepted: Dec. 13, 2005